Phylogeny and genomics of SAUL, an enigmatic bacterial lineage frequently associated with marine sponges

Carmen Astudillo-García[1,2], Beate M. Slaby[3,4], David W. Waite[5], Kristina Bayer[3], Ute Hentschel[3,6], Michael W. Taylor[1,7]*

[1]School of Biological Sciences, University of Auckland, Auckland, New Zealand

[2]Institute of Marine Science, University of Auckland, Auckland, New Zealand

[3]RD3 Marine Microbiology, GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

[4]Department of Botany II, Julius-von-Sachs Institute for Biological Sciences, University of Würzburg, Würzburg, Germany

[5]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD, Australia

[6]Christian-Albrechts-Universität zu Kiel, Kiel, Germany

[7]Maurice Wilkins Centre for Molecular Biodiscovery, University of Auckland, New Zealand

*Corresponding author:

Associate Professor Michael W. Taylor

School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

Email: mw.taylor@auckland.ac.nz; Telephone +64 9 9232280

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1111/1462-2920.13965

Running title: Phylogeny and genomics of SAUL

Keywords: marine sponge; symbiont; phylogeny; metagenome binning; functional analyses

ORIGINALITY-SIGNIFICANCE STATEMENT

Marine sponges contain an extraordinary level of microbial diversity, mostly comprised of uncultured microorganisms. While single-cell genomics and metagenomics approaches have recovered genomes from certain sponge-associated microbes, there remain some key lineages of abundant sponge symbionts about which little is known. For one such lineage that is commonly reported from sponges, the so-called "sponge-associated unclassified lineage" (SAUL), we carried out a meta-analysis of existing 16S rRNA gene datasets, coupled with extensive phylogenetic and genomic analyses, to gain new insights into the prevalence, identity and potential function of this enigmatic bacterial clade.

SUMMARY

Many marine sponges contain dense and diverse communities of associated microorganisms. Members of the "sponge-associated unclassified lineage" (SAUL) are frequently recorded from sponges, yet little is known about these bacteria. Here we investigated the distribution and phylogenetic status of SAUL. A meta-analysis of the available literature revealed the widespread distribution of this clade and its association with taxonomically varied sponge hosts. Phylogenetic analyses, conducted using both 16S rRNA gene-based phylogeny and concatenated marker protein sequences, revealed that SAUL is a sister clade of the candidate phylum "*Latescibacteria*". Furthermore, we conducted a comprehensive analysis of two draft genomes assembled from sponge metagenomes, revealing novel insights into the physiology of this symbiont. Metabolic reconstruction suggested that SAUL members are aerobic bacteria with facultative anaerobic metabolism, with the capacity to degrade multiple sponge- and algae-derived carbohydrates. We described for the first time in a sponge symbiont the putative genomic capacity to transport phosphate into the cell and to produce and store polyphosphate granules, presumably constituting a phosphate reservoir for the sponge host in deprivation periods. Our findings suggest that the lifestyle of SAUL is symbiotic with the host sponge, and identify symbiont factors which may facilitate the establishment and maintenance of this relationship.

INTRODUCTION

Marine sponges (phylum *Porifera*) are among the most ancient of the extant metazoans (Hooper and Van Soest, 2002) and are key components of the benthos in an array of marine habitats (Bell, 2008). Many sponges also host diverse and abundant microbial communities which constitute up to 35% of total sponge biomass (Taylor *et al.*, 2007a, 2007b; Hentschel *et al.*, 2012). These symbiont communities comprise up to 41 different bacterial phyla (Thomas *et al.*, 2016), as well as many archaea, viruses and fungi (Taylor *et al.*, 2007a; Webster and Thomas, 2016). In this study, the terms "symbiont" and "symbiosis" are used in a broad definition, to refer simply to the long-term association of two or more organisms, irrespective of the nature of this relationship, following the early de Bary definition (de Bary, 1879; Taylor *et al.*, 2007a).

The recalcitrance of many, or even most, sponge-associated microorganisms to grow in a pure laboratory culture has constrained our ability to understand the physiology of sponge symbionts. Moreover, studies of the sponge microbiota have commonly focused on bacteria that are numerically dominant and/or known to play significant functional roles, such as the cyanobacterium "*Ca*. Synechococcus spongiarum" (Erwin and Thacker, 2008; Gao *et al.*, 2014; Burgsdorf *et al.*, 2015) or the candidate phylum "*Poribacteria*" (Fieseler *et al.*, 2004; Siegl *et al.*, 2011; Kamke *et al.*, 2014). Consequently, other, less prominent members of the so-called "microbial dark matter" (Rinke *et al.*, 2013) remain poorly understood. One such clade is the sponge-associated unclassified lineage (SAUL) (Schmitt *et al.*, 2012), initially designated as PAUC34f (Hentschel *et al.*, 2002),. This clade was first identified as a symbiont of the tropical sponge *Theonella swinhoei* (Hentschel *et al.*, 2002), with subsequent studies revealing its presence in numerous sponge species (Taylor *et al.*, 2007a; Schmitt *et al.*, 2012; Simister *et al.*, 2012a; Thomas *et al.*, 2016). For example, SAUL represented approximately 12% and 6% of sequences derived from the sponges *Rhopaloeides odorabile* (Simister *et al.*, 2012) and *Ancorina alata* (Simister *et al.*, 2013), respectively. A more recent study of 81

different sponge species, comprising the Sponge Microbiome Project, found SAUL (labelled as PAUC34f) in 72 of those species (Thomas *et al.*, 2016). This apparent affinity for sponges was reflected in the assignment of >70% of SAUL sequences to so-called 'sponge-specific clusters' (Simister *et al.*, 2012a), which represent clades of microorganisms that can be highly enriched in marine sponges (Hentschel *et al.*, 2002; Simister *et al.*, 2012a). It can also be vertically transmitted, with SAUL-affiliated sequences being identified in samples from adult, embryo and larval stages of the oviparous sponge *Ectyoplasia ferox* (Gloeckner *et al.*, 2013). Vertical microbial transmission enables marine sponges to transfer relevant symbionts from adults to offspring and consequently to maintain, over time, complex and host-specific microbial communities (Schmitt *et al.*, 2012). Despite this sponge affinity, SAUL has also been detected in other environments such as seawater, marine sediments and soils (Taylor *et al.*, 2007a, 2013; Thomas *et al.*, 2016), albeit at lower abundance than in sponges.

Although the SAUL lineage is commonly associated with sponges, to date no studies have examined its relationship with the host sponge and the precise phylogenetic classification of SAUL remains unresolved. Initially classified as a member of *Deltaproteobacteria* (Hentschel *et al.*, 2002), subsequent studies have assigned SAUL-affiliated sequences as part of either *Acidobacteria* or *Deferribacteres* (Webster *et al.*, 2011), as an independent clade closely related to the *Planctomycetes-Verrucomicrobia-Chlamydiae* (PVC) superphylum (Wagner and Horn, 2006; Taylor *et al.*, 2007a; Kamke *et al.*, 2010; Simister *et al.*, 2012a), or were unable to classify such sequences beyond membership in the domain *Bacteria* (Montalvo and Hill, 2011). This lack of agreement regarding SAUL classification, together with a lack of knowledge about its genomic capabilities, motivated us to take a detailed look at this enigmatic clade. Genomics studies of individual lineages have revealed novel insights into the lifestyles of other sponge symbionts, including the candidate phylum "*Poribacteria*" (Fieseler *et al.*, 2006; Siegl *et al.*, 2011; Kamke *et al.*, 2013, 2014), the candidate genus "*Candidatus* Entotheonella" (Lackner *et al.*, 2017), the widespread cyanobacterium "*Ca.* Synechococcus spongiarum" (Tian *et al.*, 2014; Burgsdorf *et al.*, 2015) and a sponge-associated

sulphur-oxidizing *Gammaproteobacterium* (Tian *et al.*, 2017). Similar findings were obtained by metagenomic analyses of the sponge microbiota, thus revealing common features of sponge symbionts that include an enrichment of proteins involved in microbe-host signalling (Thomas *et al.*, 2010; Fan *et al.*, 2012), universal stress proteins such as UspA (Fan *et al.*, 2012) and an abundance of bacterial defence systems including CRISPR-Cas, restriction-modification and toxin-antitoxin systems (Fan *et al.*, 2012; Horn *et al.*, 2016; Slaby *et al.*, 2017).

In this study we aimed to (i) comprehensively describe the distribution and abundance of the SAUL clade, across different environments and among different sponge species, using a meta-analysis of available 16S rRNA gene sequences; (ii) use sequences derived from SAUL and related bacterial phyla to robustly infer its phylogenetic position amongst bacteria; (iii) determine the genomic potential and identify symbiotic features of SAUL members by reconstructing genomes from sponge-derived metagenome data.

RESULTS AND DISCUSSION

SAUL is widespread and abundant in marine sponge hosts

In the absence of a comprehensive evaluation of SAUL occurrence, we performed a meta-analysis to examine those sponge studies which reported the presence of SAUL. The SAUL clade was identified in 15 studies (including the recent Sponge Microbiome Project (Thomas *et al.*, 2016)) (Supporting Information Table S1) and in 93 different sponge species, with relative sequence abundances per sponge species ranging from 20.7% to less than 0.001% (Figure 1, Supporting Information Figure S1). The global SAUL distribution in sponges was wide-ranging, with its members identified from the

Pacific, Atlantic and Indian oceans, as well as the Red, Mediterranean and Caribbean seas (data not shown). The apparent high abundance of this clade, its broad distribution, and its presence in taxonomically diverse sponge species suggest a high degree of generalism within marine sponges and highlight the symbiotic potential and likely importance of SAUL in the different sponge-associated microbial communities. This prevalence is in agreement with several studies that have identified overlapping microbial community members (including SAUL) in geographically separated and phylogenetically distant sponge species (Hentschel *et al.*, 2002; Simister *et al.*, 2012a). SAUL was also recorded in seawater and sediment samples in the study by Thomas and colleagues (2016) (Figure 1B).

SAUL is a sister clade of "*Latescibacteria*"

Having demonstrated its widespread presence and numerical abundance in the sponge microbiota, we sought to determine the phylogenetic position of the SAUL lineage. Phylogenetic inference of near full-length (>1450 bp) 16S rRNA gene sequences and concatenated marker protein sequences supported the clustering of SAUL as a monophyletic clade (Figure 2). Phylogenomic analysis of up to 37 markers suggested that SAUL is a sister clade of the candidate phylum "*Latescibacteria*" (formerly WS3, Rinke *et al.*, 2013) (Figure 2A). This relationship was also observed for 16S rRNA gene sequence data, but was not supported by bootstrap resampling (Figure 2B). "*Latescibacteria"* represents a monophyletic cluster closely related to the *Fibrobacteres-Chlorobi-Bacteroidetes* (FCB) group (Rinke *et al.*, 2013; Anantharaman *et al.*, 2016; Hug *et al.*, 2016). In our analyses, both SAUL and "*Latescibacteria*" clustered together with *Fibrobacteres*, *Chlorobi* and *Bacteroidetes*, albeit with weaker bootstrap support (75%), and previous research has also demonstrated that these five phyla are not monophyletic (Hug *et al.*, 2016). With the aim of revealing how SAUL is related at the genomic level to *"Latescibacteria"*, we investigated genomic similarities between these lineages (Supporting Text). Low genomic similarity was observed between SAUL and "*Latescibacteria*", likely

reflecting different lifestyles related to the disparate environments with which they are associated ("*Latescibacteria*" members are typically free-living bacteria found in terrestrial, aquatic and marine environments (Rinke *et al.*, 2013; Youssef *et al.*, 2015)).

Inconsistency between 16S rRNA gene-based phylogeny and phylogenomic analysis was observed previously for the sponge-associated clade "*Poribacteria*" (Kamke *et al.*, 2014), which also clustered SAUL sequences as a monophyletic sister clade to "*Latescibacteria"*. While a concatenation of different marker protein sequences can provide higher resolution for resolving intra- and inter-phylum level relationships compared with analysis of a single marker gene such as 16S rRNA gene, such approaches are limited by a small number of available draft genomes. In our study, only three draft genomes with sufficient completeness to be used for phylogenomic tree reconstruction were available for each of SAUL and "*Latescibacteria"*.

To further evaluate the phylogenetic status of SAUL, we calculated the average 16S rRNA gene sequence similarity within SAUL and between members of SAUL and those of other bacterial phyla (Supporting Information Table S2). Average 16S rRNA gene sequence similarity within the SAUL cluster was 88%, and its average similarity with "*Latescibacteria"*, its closest relative according to phylogenomic analyses, was 80.8%. According to recently suggested threshold sequence criteria for phylum, class and order levels (75%, 78.5% and 82%, respectively; Yarza *et al.*, 2014), SAUL and "*Latescibacteria"* would represent sister clades, possibly different classes within the same phylum. Although our results revealed that SAUL is a lineage closely related with the FCB superphylum, and it is reproducibly a sister clade of "*Latescibacteria*", the paucity of near-complete genomes for SAUL and other closely related clades prevent further assertions from being made confidently. As a consequence, the phylogenetic status of the SAUL clade must be revisited once more genomic data are available.

Internal structure of the SAUL clade

In addition to determining the placement of SAUL within the bacterial tree of life, we sought to characterise the internal phylogenetic structure of this lineage. Phylogenetic trees based on 16S rRNA gene sequences showed the existence of three subgroups of SAUL sequences clustering independently from each other (Figure 3). High bootstrap scores (>80%) supported the branching for those three clusters, hereafter referred to as Clusters I, II and III. The application of taxonomic thresholds based on 16S rRNA gene sequence similarity to interpret the internal architecture of the SAUL clade suggested these clusters may represent distinct families (Supporting Information Table S2). Most SAUL unique representative sequences are sponge derived (70.8%), with seawater, sediment and other marine source-derived sequences comprising a smaller fraction (24.7%). Only 4.5% of SAUL-affiliated sequences were derived from non-marine sources, primarily freshwater or biofilms. Origins varied when evaluating the three clusters individually, with 62.2%, 87.5% and 42% of the sequences being derived from marine sponges for Clusters I, II and III, respectively. The cluster with the highest representation of sponge-derived sequences (Cluster II) contains the first identified SAUL sequence (clone PAUC34f, AF186412). A 16S rRNA gene sequence derived from one of the SAUL metagenome bins described below is also included in the same cluster (bin_petrosia). The 16S rRNA gene sequences derived from the two other SAUL bins (bin_aplysina and bin_aplysina_2) are in Cluster I.

Functional potential and symbiont characteristics of SAUL revealed by population genome binning

Assembly of metagenome data from *Aplysina aerophoba* and *Petrosia ficiformis* led to the reconstruction of two near-complete draft genomes, with 90.3% and 86.5% completeness (based on the identification of 104 markers (Parks *et al.*, 2015)), and an estimated genome size of 6.3 and 4.7 Mbp for bin_aplysina and bin_petrosia, respectively (Table 1). A third bin (bin_aplysina_2), also

constructed from *Aplysina aerophoba* metagenomic data, was not used for further genomic analyses due to low completeness (39.42%). Functional annotation of 4,932 (bin_aplysina) and 3,711 (bin_petrosia) genes enabled a comprehensive analysis of the metabolic potential and biosynthetic capabilities of SAUL (Figure 4). It is important to note that, due to genome incompleteness, any apparent lack of specific enzymes/proteins should be interpreted with caution. Metabolic reconstruction of the two metagenome bins suggested that SAUL members are aerobic bacteria with facultative anaerobic metabolism, possessing also the capacity to degrade multiple sponge- and algae-derived carbohydrates. Genes involved in major central pathways, such as the tricarboxylic acid cycle, glycolysis, pentose phosphate pathway, Wood-Ljungdahl pathway and oxidative phosphorylation, were identified in at least one of the bins. Moreover, genes encoding several enzymes involved in the uptake and/or metabolism of nitrogen and sulphate were identified in either one or both SAUL bins. We also detected genes involved in phosphate transport and metabolism, including enzymes encoding the high affinity phosphate transporter and control of PHO regulon (Figure 4, Box (A)), as well as the enzyme polyphosphate kinase (*ppk*, EC 2.7.4.1), suggesting that this clade may be involved in phosphorus sequestration from the environment and its later conversion into the polyphosphate (polyP) storage form. Although the presence of polyP granules in bacterial cells has been described previously in the associated communities of three phylogenetically divergent sponge species (Zhang *et al.*, 2015), this is the first time that the genomic potential for polyP granules production has been identified in an actual sponge associate.

Both genomes encoded enzymes involved in the production of amino acids, vitamins, purines and pyrimidines, as well as near-complete replication, transcriptional and translational machineries. Additionally, the genomic machinery to release and conserve energy via the electron transport chain and oxidative phosphorylation, as well as substrate-level phosphorylation, was revealed for both bins (more detailed discussion of specific aspects of central metabolism, biosynthesis and information transfer machinery in SAUL genomes can be found in Supporting Text).

Secondary metabolite biosynthesis

Production of biologically active secondary metabolites is an important defence mechanism utilised by sponges for protection against predators or epibionts (Pawlik, 2011). Many secondary metabolites are produced by polyketide synthases (PKS), mainly Type I PKS, and non-ribosomal peptide synthetases (NRPS) (Fischbach and Walsh, 2006; Newman and Cragg, 2012). The origin of many compounds remains controversial (Hentschel *et al.*, 2012), with some produced by the sponge and others by associated microorganisms (Piel *et al.*, 2004; Taylor *et al.*, 2007a; Sala *et al.*, 2014; Wilson *et al.*, 2014; Flórez *et al.*, 2015). Both SAUL metagenome bins contained genes encoding for PKS modules and related proteins (COG3321). Further analysis with antiSMASH (Weber *et al.*, 2015) revealed the presence of several secondary metabolite biosynthesis gene clusters. In both SAUL genomes, Type I PKSs were identified (Figure 4, Box (B)), as well as several putative clusters. A BlastP search conducted with the PKSs identified in both bins showed >60% sequence similarity to a sponge symbiont ubiquitous Type I PKS (Sup) identified in *Theonella swinhoei* (cosmid pSW1H8) (Fieseler *et al.*, 2007). That same study identified the PKS in 10 additional sponge species, all of which, including *T. swinhoei*, belonged to the "high-microbial-abundance" group (Hentschel *et al.*, 2003), which also contains *A. aerophoba* and *P. ficiformis* (Gloeckner *et al.*, 2014). The abundance of microorganisms present in these sponges may lead to intense (and not necessarily positive) microbe-microbe interactions (Thomas *et al.*, 2016). The production of secondary metabolites with antimicrobial properties may be used as a defence strategy by some sponge symbionts to confront either other symbionts present in the community or foreign microorganisms that enter the sponge environment.

Host-microbe recognition systems through eukaryotic-like proteins

Adaptive symbiosis factors such as eukaryotic-like proteins (ELPs) in bacterial symbionts, particularly ankyrin (ANK), tetratricopeptide (TPR) and leucine-rich (LRR) repeat proteins, have attracted much attention due to their presumed involvement in mediating host-microbe recognition and interaction, improving attachment to the eukaryotic host and avoidance of the host's immune response (Habyarimana *et al.*, 2008; Al-Khodor *et al.*, 2010; Liu *et al.*, 2011; Siegl *et al.*, 2011; Fan *et al.*, 2012; Cerveny *et al.*, 2013; Reynolds and Thomas, 2016). Recent (meta)genomic studies of sponge-associated microbial communities have identified such factors as being widespread in these symbiont communities (Liu *et al.*, 2011, 2012; Fan *et al.*, 2012; Kamke *et al.*, 2014; Tian *et al.*, 2014; Burgsdorf *et al.*, 2015). Due to their ubiquity in sponge-associated microbial communities and their presumed importance in symbiont recognition, we investigated the presence of genes encoding for ELPs in SAUL bins. Screening of COG and PFAM databases revealed that both SAUL genomes encoded ANKs (annotated as COG0666, PF00023), TPRs (COG5010, PF00515, PF07719), and LRR (COG4886). Genes encoding for another ELP, WD40 repeat proteins (PF00400), were also identified in bin_aplysina. The finding of these ELPs within SAUL genomes is consistent with previous sponge microbiota studies, and suggests their postulated importance for symbiont recognition by the sponge host.

Evidence for SAUL adaptation to host conditions

Sponge symbionts exhibit resistance mechanisms designed to specifically address changes in host conditions that generally lead to stress. In this context, genomic studies of sponge-associated microbial communities have revealed an enrichment of stress-related proteins (López-Legentil *et al.*, 2008; Fan *et al.*, 2012; Liu *et al.*, 2012) that may help symbionts cope with environmental stressors, including the presence of antimicrobial compounds (Piel, 2009), bioaccumulation of heavy metals (Hansen *et al.*, 1995; Webster *et al.*, 2001), and changes in temperature (Fan *et al.*, 2013), pH (Ribes *et al.*, 2016) and sedimentation (Luter *et al.*, 2012). The SAUL clade possesses genomic signatures

related to adaptation to the microenvironment of the host sponge, with both bins carrying genes encoding proteins such as UspA (universal stress protein A) (annotated as COG0589), which is synthesised in response to environmental stress such as heat and/or osmotic shock, nutrient starvation, or exposure to heavy metals (Nyström and Neidhardt, 1994; Kvint *et al.*, 2003). Furthermore, genes encoding for the complex PotABCD, involved in the uptake of polyamines such as putrescine, spermidine (the most common polyamides in bacteria (Wortham *et al.*, 2007)), and cadaverine were identified in SAUL bins. Polyamines play an important role in acid resistance and can act as free radical ion scavengers (Wortham *et al.*, 2007). Proteins involved in elimination of denatured and/or damaged proteins were also identified in SAUL bins, including chaperone proteins GroEL (HSP60, COG0459), membrane proteases HflC (COG0330) and DnaK (COG0443). Several enzymes involved in cell defence against oxidative stress induced by the sponge host were also identified within SAUL genomes. These include alkyl hydroperoxide reductase subunit C (AphC, COG0450) and manganese superoxide dismutase (MnSOD, EC 1.15.1.1, COG0605). Only bin_aplysina encoded the enzyme glutathione peroxidase (EC 1.11.1.6, COG0386). The enzymes AphC and glutathione peroxidase reduce organic and lipid peroxides, respectively, with the former also protecting cells from reactive nitrogen intermediates (Chen *et al.*, 1998). Furthermore, MnSOD is an enzyme member of the superoxide dismutase family, which is one of the cell's major defence mechanisms against oxidative stress. These enzymes catalyse the conversion of superoxide molecules to hydrogen peroxide and molecular oxygen (McCord and Fridovich, 1969). Moreover, genomic comparisons of SAUL bins with two draft genomes of the closest relative, the free-living "*Latescibacteria*", indicated the apparent absence of UspA, as well as AphC and MnSOD, from "*Latescibacteria*", supporting the notion that these features commonly identified in sponge symbionts may be involved in adaptation to the host environment (see Supporting Text for a detailed genomic comparison).

Horizontal gene transfer and defence mechanisms

Horizontal gene transfer plays an important role in adaptation and evolution of both prokaryotes and eukaryotes (Keeling and Palmer, 2008; Wiedenbeck and Cohan, 2011). In sponge-associated microbes, adaptation to either specific niches or to changes in environmental conditions can be facilitated by mobile genetic elements such as transposons, plasmids and prophages (Fan *et al.*, 2012; Alex and Antunes, 2015). SAUL bins encoded for transposable insertion elements such as transposases (COG1943, COG3415, and COG3328), retroid elements containing reverse transcriptase (COG3344, PF00078) and integrases (PF00665).

Restriction-modification systems are considered bacterial defence systems that may facilitate horizontal DNA exchange between sponge symbionts but at the same time protect against DNA exchange with non-symbiont and/or pathogen microorganisms (Vasu and Nagaraja, 2013; Horn *et al.*, 2016; Slaby *et al.*, 2017). COGs including specific DNA modification and restriction systems were also identified in SAUL bins. These included Type I (COG0286, COG0610, COG 4096, PF02384, PF12161 and PF01420), Type II (COG0270, COG1743, COG0863, COG0338, COG4489 and PF00145) and Type III (COG2189 and PF04851) restriction-modification systems. The presence of these transposable elements within SAUL genomes likely confers upon these microorganisms the capacity for genetic exchange and rearrangement. This could allow for the acquisition of functions by the symbiont that maintain and strengthen its interaction with the host. Accordingly, a lower abundance of restriction-modification systems was found when investigating the draft genomes of "*Latescibacteria*". In this case, COGs included in Type I (COG0286) and Type II (COG0863 and COG0338) systems were identified in only one of the two "*Latescibacteria*" SAGs (WS3_E07).

To further investigate the magnitude of HGT events within SAUL genomes, we identified candidate transferred genes using HGT-Finder (Nguyen *et al.*, 2015). Very few putative transferred genes were found in SAUL genomes (7 and 5 for bin_aplysina and bin_petrosia, respectively), likely corresponding to ancestral HGT events ($R \leq 0.4$).

Due to the intense pumping activity of the host sponge, associated microbes are likely exposed to a large amount of viral particles and phages. Members of sponge microbial communities have thus incorporated into their genomes systems to effectively protect themselves and minimise the introduction of foreign DNA into their chromosomes (Fan *et al.*, 2012; Horn *et al.*, 2016). In this context, clustered, regularly interspaced, short, palindromic repeats (CRISPRs) and their associated proteins (Cas) (Makarova *et al.*, 2011) are commonly enriched in sponge-associated microbial communities (Thomas *et al.*, 2010; Fan *et al.*, 2012; Burgsdorf *et al.*, 2015; Horn *et al.*, 2016). CRISPR-Cas systems are heritable and adaptive immune systems that are encoded by most archaea and many bacteria. These systems are comprised of two main stages: the adaptation stage, involving incorporation of small fragments of foreign DNA into an array of spacer sequences within the CRISPR locus of the host genome; and the interference stage, where the recently acquired spacers are used to target and cleave invading DNA (Deveau *et al.*, 2010; Makarova *et al.*, 2011).

Screening of SAUL genomes revealed six CRISPR regions in bin_aplysina. Of these, two contained CRISPR-associated (Cas) proteins, thus forming two CRISPR-Cas systems: NODE_2846 and NODE_3759 (Supporting Information Table S3). NODE_2846 presents a region of 7304 nucleotides consisting of 101 spacer regions separated by a repeat of 37 nt. Cas proteins were found upstream of the CRISPR region, and included the universal Cas1 and Cas2, as well as other associated proteins characteristic of subtype I-C (Supporting Information Table S3a). Moreover, NODE_3759 has a region of 4173 nt consisting of 68 spacer regions separated by a 29 nt repeat. Apart from the universal Cas1 and Cas2, this system included proteins characteristic of subtype I-E (Supporting Information Table S3b). By contrast, bin_petrosia only contained one confirmed CRISPR region and no Cas proteins were identified. As a consequence, the functionality of this system could not be assessed. Pairwise comparisons of CRISPR spacer regions identified in SAUL genomes indicated that the SAUL members representing each bin are exposed to different types of foreign DNA. Potential targets of the spacers were mainly unknown targets, although six and one spacers from NODE_2846 and NODE_3759, respectively, registered hits in plasmids. No spacer had hits in known phages or

viruses. Moreover, no confirmed CRISPR regions were identified in any of the "*Latescibacteria*" genomes, suggesting a lower exposure to potential invading DNA in their environment.

SAUL has the potential to degrade sponge- and algae-derived carbohydrates

Both SAUL bins present multiple enzymes involved in the utilisation of diverse carbon sources (see Supporting Text for more detailed information on dedicated sugar catabolic pathways). To further evaluate SAUL's putative capacity for carbohydrate degradation, SAUL genomes were screened for carbohydrate-active enzymes (CAZymes) using the CAZY database in the web server dbCAN (Yin *et al.*, 2012). SAUL bins were rich in genes encoding glycoside hydrolases (GH), glycoside transferases (GT) and, to a lesser extent, polysaccharide lyases (PL) and carbohydrate esterases (CE) (Supporting Information Table S4). Overall, 24 different GH families were detected in SAUL bins (Supporting Information Table S5). The most abundant GH family identified was GH109, the activity of which has been described as α-N-acetylgalactosaminidase (EC 3.2.1.49). Physiological substrates for this enzyme include glycolipids, glycopeptides and glycoproteins, compounds typically found within the sponge mesohyl and as dissolved organic matter in seawater (Genin *et al.*, 2004; Blunt *et al.*, 2017). Proteins assigned to this family in SAUL bins were mostly annotated as myo-inositol 2-dehydrogenase, oxidoreductases or as predicted dehydrogenases and related proteins. Family GH33 was the second most abundant family in both SAUL bins. SAUL proteins in this family were annotated as sialidase (EC 3.2.1.18), an enzyme that hydrolyses glycosidic linkages of terminal sialic acid residues, which are present in marine sponges (Garrone *et al.*, 1971). Also present in SAUL bins, albeit to a lesser extent, was the family GH113, which contains the enzyme β-mannanase (EC 3.2.1.78). This enzyme hydrolyses the (1->4)-beta-D-mannosidic linkages in the storage plant polysaccharides mannans, galactomannans and glucomannans. Mannan replaces cellulose as the principal component of the cell wall skeleton in certain species of algae (Frei and Preston, 1961, 1964), and forms microfibrils in green algae such as *Codium fragile* and *Acetabularia crenulata*

(Mackie and Preston, 1968). Moreover, members of GH families involved in cellulose degradation, such as cellulases (endoglucanases, EC 3.2.1.4, GH5) and beta-glucosidases (GH116), were also identified in SAUL bins. These algae-derived compounds may be made available for SAUL utilisation either by the sponge host taking the compounds up directly from the surrounding seawater or as a by-product of the sponge feeding on algae. Similarly, genomic analyses of the widespread sponge symbiont "*Poribacteria*" revealed a complex suite of genes related to the degradation of several carbohydrates (Kamke *et al.*, 2013).

Concluding remarks

We have demonstrated here that the SAUL lineage is widespread, and often abundant, in high microbial abundance (HMA) sponge hosts, though can also occur in lower numbers in low microbial abundance (LMA) sponges and other non-sponge habitats. The available data collected here set the SAUL lineage close to the FCB superphylum and as a sister clade of the candidate phylum "*Latescibacteria*". However, the paucity of near-complete genomes for SAUL and other closely related clades prevents further assertions from being made confidently. Extensive genomic analyses revealed genomic characteristics that are commonly described for sponge-associated microorganisms, which may facilitate establishment and maintenance of the symbiotic relationship. These symbiosis factors include an apparent abundance of ELPs, universal stress proteins and defence mechanisms such as CRISPR-Cas.

EXPERIMENTAL PROCEDURES

Meta-analysis of available SAUL 16S rRNA gene sequences

The meta-analysis took into account those studies published up to July 2016 on the sponge microbiome in which SAUL and/or PAUC34f were identified and explicitly mentioned (Supporting

Information Table S1). Where possible, the relative sequence abundance of SAUL in different sponge species was noted for each study. When this information was not available, sequence data were downloaded and relative abundance was calculated as percentage of sequences assigned to either SAUL or PAUC34f per sponge species.

Deciphering SAUL phylogeny

Long 16S rRNA gene sequences (≥1200 bp) previously classified as being affiliated with SAUL (Simister *et al.*, 2012a) were used as reference sequences to conduct an extensive BLAST search against the GenBank nr/nt database. The 100 best hits with >85% sequence identity for each search were retained, and a phylogenetic tree was constructed to confirm the SAUL affiliation of the selected sequences. As of May 2016, all SAUL-affiliated 16S rRNA gene sequences available in GenBank were retrieved and included in our analyses, as well as two 16S rRNA gene sequences derived from two SAUL metagenomes (see below). Accession numbers for long SAUL sequences utilised in the study are listed in Supporting Information Table S6. SAUL sequences were aligned using the SINA Web Aligner (Pruesse *et al.*, 2007), merged with the SILVA 119 SSU Ref NR 99 database, and imported into ARB for further manual curation of the alignment. Phylogenetic analyses were carried out on near-full length (≥1450 bp) SAUL 16S rRNA gene sequences together with sequences representing closely related phyla (as per (Simister *et al.*, 2012a)). Shorter sequences were subsequently added without changing tree topology using the Parsimony Interactive tool in ARB. Trees were constructed in ARB using neighbour-joining (Jukes-Cantor correction) and maximum likelihood (RAxML) to assess the robustness of the constructed phylogeny. Maximum likelihood trees were constructed using three different distribution models, GTRMIX (default), GTRGAMMA and GTRCAT. The outgroup for tree calculation comprised sequences belonging to the distantly related clades *Thermotogae* and *Aquificae*. Sequence conservation filters

(50%) were applied for tree construction (Ludwig *et al.*, 1998), and bootstrap analyses were done with 500 resamplings.

To investigate sub-clusters within the SAUL lineage, we selected long SAUL-affiliated 16S rRNA gene sequences (≥1450 bp) and a range of sequences from several bacterial phyla as outgroup (*Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*, *Lentisphaerae*, "*Poribacteria*", *Latescibacteria* (WS3), candidate division OP3, *Firmicutes* and candidate division BRC1). Phylogenetic trees were constructed using the same methodology described earlier for 16S rRNA gene-based phylogeny, except that a conservation filter was not applied in order to include the maximum number of sequence alignment positions in the analysis.

Average sequence similarity among clades has been used, in addition to phylogenetic tree construction, to help decipher the phylogeny of a given microorganism (Yarza *et al.*, 2014). Thus, 16S rRNA gene sequence similarity within the SAUL clade, and between SAUL and other clades considered in the phylogenetic analysis, was calculated by applying the similarity option of the ARB Distance Matrix tool.

A custom data set of 37 different marker protein sequences (Supporting Information Table S7) was used to conduct a phylogenomic analysis (Rinke *et al.*, 2013). SAUL gene sequences employed for this analysis were obtained from three unpublished draft genomes: two of these (bin_petrosia and bin_aplysina) are analysed in detail in this study (see Results and Discussion section), while the third was insufficiently complete for full genome analysis. Each marker gene was identified, requiring >30% coverage of the protein sequence and e-value <0.001. Where multiple homologues were identified in a single genome, only the best match was retained. Homologues were then aligned to their respective reference alignment using HMMER (v3.1b2). Alignments were cleaned with Gblocks (Castresana, 2000) and the markers were concatenated. A tree was then built in RAxML, using the WAG+Gamma model, and bootstrapping was calculated with 100 resamplings.

Sponge sample collection and metagenome sequencing

Samples of the Mediterranean sponges *Petrosia ficiformis* and *Aplysina aerophoba* were collected and their metagenomes were sequenced and assembled for previous studies (Horn *et al.*, 2016; Slaby *et al.*, 2017). Raw sequence reads obtained from *P. ficiformis* were inspected using FastQC 0.11.2 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for adapters, overall quality, length and ambiguous bases. Reads were trimmed with Trimmomatic 0.31 (PR –phred 33 LEADING:3 ILLUMINACLIP:2:30:10) (Bolger *et al.*, 2014) then merged using bbmerge (https://sourceforge.net/projects/bbmap/). Merged and unmerged reads were again subjected to Trimmomatic for further quality trimming and length filtering (SE –phred 33 SLIDINGWINDOW:4:25 MINLENGTH:150 AVGQUAL:30). The remaining reads were assembled with IDBA-UD 1.1.1 (-mink 10 maxk -100) (Peng *et al.*, 2012). Contigs shorter than 1000 bp were discarded.

The metagenomes from *A. aerophoba* and *P. ficiformis* were binned using the software CONCOCT v. 0.4.0 at default settings (Alneberg *et al.*, 2014), with preparation of the coverage tables for the binning process as described in Slaby *et al.* (2017). A fasta file for each bin was created with the in-house python script mkBinFasta.py (https://github.com/bslaby/scripts/). The identification of rRNA genes was conducted with nhmmer (Wheeler and Eddy, 2013). Bin completeness was estimated by conducting an hmmsearch against a database of 104 essential genes using CheckM (Parks *et al.*, 2015). Target bins were identified by a BLASTn search of 86 known SAUL 16S rRNA gene sequences against a BLAST database of the rRNA genes identified in the metagenomic bins, with an identity cut-off of 85%. To confirm their affiliation to the SAUL clade, bin-derived sequences were then used to construct a phylogenetic tree with previously identified SAUL sequences. The identified SAUL bins bin_aplysina and bin_aplysina_2 (both derived from *A. aerophoba*), and bin_petrosia, were refined manually *via* a previously published R pipeline (Albertsen *et al.*, 2013) and contigs shorter than 2000 bp were filtered out. Raw Illumina data for *A. aerophoba* are deposited under JGI's GOLD study ID

Gs0099546. The assembly data are deposited in GenBank under the accession number MKWU00000000. Sequence data for *P. ficiformis* are deposited in the Sequence Read Archive (SRA) under the BioProject PRJNA318959 and the BioSample SAMN04870510 (SRA: SRP074318, WGS:LXNJ00000000).

The draft genomes of the two most complete SAUL bins (bin_aplysina and bin_petrosia) were then submitted to RAST, the SEED-based prokaryotic genome annotation server (Aziz *et al.*, 2008; Overbeek *et al.*, 2014), for automated open reading frames (ORF) prediction and annotation of SEED subsystems, followed by manual checking. Clusters of orthologous groups (COGs) (Tatusov *et al.*, 2003) were annotated using rpsBLAST (v. 2.2.15), while Pfam (Finn *et al.*, 2016) and TIGRfam (Haft *et al.*, 2003) protein families were identified with HMMER 3.0. All annotations were conducted through the WebMGA (Wu *et al.*, 2011) function annotation tool, with an e-value cut-off of 0.001. Additionally, SAUL predicted genes were submitted to GhostKOALA automatic annotation server for KEGG (Kanehisa *et al.*, 2004) annotation by GHOSTX searches against a non-redundant set of KEGG genes (Kanehisa *et al.*, 2016). Carbohydrate active enzymes (CAZymes) (http://www.cazy.org) were identified by searching translated protein sequences against dbCAN HMMs (Yin *et al.*, 2012) using HMMER 3, and results were filtered using an e-value cut-off of 0.00001. Additionally, all CAZy hits were manually evaluated with SEED annotation and excluded when results were conflicting. SAUL genomes were searched using blastn (v +2.6.0) against the GenBank NR database. The BLAST output was subjected to HGT-Finder (*R* threshold ranging from 0.2 to 0.9, *Q* value <0.05) to identify HGT candidates (Nguyen *et al.*, 2015). Clustered regularly interspaced short palindromic repeats (CRISPR) were identified using CRISPRFinder (Grissa *et al.*, 2008) in CRISPRs web server with default settings. Both confirmed and candidate CRISPRs were identified, but only confirmed CRISPR regions were used for further analysis. Furthermore, CRISPR-associated proteins (Cas) were identified using SEED annotation, and classified as described previously (Makarova *et al.*, 2015). Pairwise comparisons of the confirmed CRISPR spacer sequences were conducted with CRISPRcompar (Grissa *et al.*, 2008). Putative targets of spacers were identified with CRISPRTarget (Biswas *et al.*, 2013) using GenBank-

Phage, GenBank-Plasmid, ACLAME and RefSeq-Viral databases (default setting except: gap open -5, e-value:0.1, cut-off score:20). With the aim of exploring the putative capability of SAUL to produce secondary metabolites, both bins were analysed with the web-based tool antiSMASH (version 3.0) (Weber *et al.*, 2015) for the detection of secondary metabolite biosynthetic gene clusters.

*SAUL-"*Latescibacteria*" genome comparison*

SAUL draft genomes were compared to those of the closest related phylum, the candidate phylum "*Latescibacteria*". Two "*Latescibacteria*" SAGs, WS3_E07 and WS3_B13, were previously reconstructed and analysed by Rinke and colleagues (2013). These two SAGs were downloaded from GenBank (assembly IDs: NZ_AQSL00000000.1 and ASWY00000000.1 respectively) and submitted to the RAST web server for ORF prediction and annotation. Sequence-based comparison between SAUL bins and the two "*Latescibacteria*" SAGs was then conducted using RAST comparative tools. Additionally, information on COG annotation obtained from Integrated Microbial Genomics (IMG) (SAG ID: SCGC AAA252-B13 and SCGC AAA252-E07) was used for further comparisons between SAUL bins and "*Latescibacteria*" SAGs using STAMP V2.1.3 (Parks *et al.*, 2014).

ACKNOWLEDGEMENTS

CAG was supported by an Encouraging and Supporting Innovation Doctoral Scholarship in Marine Science awarded by the University of Auckland. BMS was supported by a grant of the German Excellence Initiative to the Graduate School of Life Sciences, University of Würzburg, and by the European Union's Horizon 2020 research and innovation program under Grant Agreement no. 679849 ('SponGES'). We thank Hannes Horn for his contribution to processing of the metagenomic data of *Petrosia ficiformis* and valuable discussions.

LIST OF FIGURE AND TABLE LEGENDS

Figure 1: Relative abundance (percentage) of SAUL- or PAUC34f-affiliated sequences in 16S rRNA gene-based studies of marine sponge bacterial communities. (A) Prevalence of SAUL in different sponge species across 14 sponge microbiota studies. Letters beside bars represent sponge species belonging to the same study defined in Table S1. (B) Prevalence of SAUL in 72 of the 81 sponge species reported in the recent Sponge Microbiome Project study (Thomas *et al.*, 2016). Where known, sponge species are classified as high (black bars) or low microbial abundance (white bars), according mainly to according mainly to Gloeckner *et al.* (2014) and Moitinho-Silva *et al.* (2017).

Figure 2: Phylogeny showing position of the SAUL clade relative to other bacterial phyla. (A) Phylogenomic analysis based on a data set of up to 37 concatenated marker proteins. (B) 16S rRNA gene sequence-based maximum-likelihood analysis of SAUL and its closest relatives. Bootstrap support (100 resamplings for (A), 500 for (B)) is shown on tree nodes where support is either ≥90% (filled circles) or ≥75% (open circles). Scale bars represent 10% sequence divergence.

Figure 3. 16S rRNA gene-based maximum likelihood phylogenetic analysis showing the internal architecture of the SAUL clade. Bold sequences denote sequences derived from marine sponges; black arrows indicate sequences obtained from SAUL bins (bin_aplysina, bin_aplysina_2 and bin_petrosia) as well as the first SAUL/PAUC34f sequence identified (AF186412). Details are the same as those provided for Figure 2A.

Figure 4. Schematic overview showing genomic potential of a SAUL cell. Black lines indicate functions identified in both metagenome bins. Red and green lines/transporters indicate functions only present in either bin_aplysina or bin_petrosia, respectively. Dashed lines indicate that some enzymes of the pathway were not identified in the bins (see Supporting Text for details). (A) Model

of phosphate operon signal transduction; (B) Features of the Type I polyketide synthase identified in bin_aplysina and bin_petrosia. Abbreviations of PKS domains as follows: KR, ketoreductase; KS, ketosynthase; AT, acyltransferase; ACPS, acyl carrier protein; ER, Enoyl reductase.

Table 1. General information about SAUL metagenome bins.

REFERENCES

Al-Khodor, S., Price, C.T., Kalia, A., and Abu Kwaik, Y. (2010) Ankyrin-repeat containing proteins of microbes: a conserved structure with functional diversity. *Trends Microbiol.* **18**: 132–139.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**: 533–538.

Alex, A. and Antunes, A. (2015) Whole genome sequencing of the symbiont *Pseudovibrio* sp. from the intertidal marine sponge *Polymastia penicillus* revealed a gene repertoire for host-switching permissive lifestyle. *Genome Biol. Evol.* **7**: 3022–3032.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., et al. (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**: 1144–1146.

Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**: 7:13219. doi:10.1038/ncomms13219.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., et al. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

de Bary, A. (1879) Die erscheinung der symbiose. Verlag von Karl J. Trübner, Strassburg.

Bell, J.J. (2008) The functional roles of marine sponges. *Estuar. Coast. Shelf Sci.* **79**: 341–353.

Biswas, A., Gagnon, J.N., Brouns, S.J.J., Fineran, P.C., and Brown, C.M. (2013) CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol.* **10**: 817–827.

Blunt, J.W., Copp, B.R., Keyzers, R.A., Munro, M.H.G., and Prinsep, M.R. (2017) Marine natural products. *Nat. Prod. Rep.* **34**: 235–294.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence

data. *Bioinformatics* **30**: 2114–2120.

Burgsdorf, I., Slaby, B.M., Handley, K.M., Haber, M., Blom, J., Marshall, C.W., et al. (2015) Lifestyle evolution in cyanobacterial symbionts of sponges. *MBio* **6**: e00391–15. doi:10.1128/mBio.00391–15.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**: 540–552.

Cerveny, L., Straskova, A., Dankova, V., Hartlova, A., Ceckova, M., Staud, F., and Stulik, J. (2013) Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms. *Infect. Immun.* **81**: 629–635.

Chen, L., Xie, Q.W., and Nathan, C. (1998) Alkyl hydroperoxide reductase subunit C (AhpC) protects bacterial and human cells against reactive nitrogen intermediates. *Mol Cell* **1**: 795–805.

Deveau, H., Garneau, J.E., and Moineau, S. (2010) CRISPR/Cas system and its role in phage-bacteria interactions. *Annu. Rev. Microbiol.* **64**: 475–493.

Erwin, P.M. and Thacker, R.W. (2008) Phototrophic nutrition and symbiont diversity of two Caribbean sponge-cyanobacteria symbioses. *Mar. Ecol. Prog. Ser.* **362**: 139–147.

Fan, L., Liu, M., Simister, R., Webster, N.S., and Thomas, T. (2013) Marine microbial symbiosis heats up: the phylogenetic and functional response of a sponge holobiont to thermal stress. *ISME J.* **7**: 991–1002.

Fan, L., Reynolds, D., Liu, M., Stark, M., Kjelleberg, S., Webster, N.S., and Thomas, T. (2012) Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proc. Natl. Acad. Sci. USA* **109**: 1878–1887.

Fieseler, L., Hentschel, U., Grozdanov, L., Schirmer, A., Wen, G., Platzer, M., et al. (2007) Widespread occurrence and genomic context of unusually small polyketide synthase genes in microbial

consortia associated with marine sponges. *Appl. Environ. Microbiol.* **73**: 2144–2155.

Fieseler, L., Horn, M., Wagner, M., and Hentschel, U. (2004) Discovery of the novel candidate phylum "*Poribacteria*" in marine sponges. *Appl. Environ. Microbiol.* **70**: 3724–3732.

Fieseler, L., Quaiser, A., Schleper, C., and Hentschel, U. (2006) Analysis of the first genome fragment from the marine sponge-associated, novel candidate phylum *Poribacteria* by environmental genomics. *Environ. Microbiol.* **8**: 612–624.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., et al. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**: 279–285.

Fischbach, M.A. and Walsh, C.T. (2006) Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic machinery, and mechanisms. *Chem. Rev.* **106**: 3468–3496.

Flórez, L. V, Biedermann, P.H.W., Engl, T., and Kaltenpoth, M. (2015) Defensive symbioses of animals with prokaryotic and eukaryotic microorganisms. *Nat. Prod. Rep.* **32**: 904–36.

Frei, E. and Preston, R.D. (1964) Non-cellulosic structural polysaccharides in algal cell walls. II. Association of xylan and mannan in *Porphyra umbilicalis*. *Proc. R. Soc. B* **160**: 314–317.

Frei, E. and Preston, R.D. (1961) Variants in the structural polysaccharides of algal cell walls. *Nature* **192**: 939–943.

Gao, Z.M., Wang, Y., Tian, R.M., Wong, Y.H., Batang, Z.B., Al-Suwailem, A.M., et al. (2014) Symbiotic adaptation drives genome streamlining of the cyanobacterial sponge symbiont "Candidatus *Synechococcus spongiarum*." *MBio* **5**: e00079–14. doi:10.1128/mBio.00079–14.

Garrone, R., Thiney, Y., and Pavans de Ceccatty, M. (1971) Electron microscopy of a mucopolysaccharide cell coat in sponges. *Experientia* **27**: 1324–1326.

Genin, E., Njinkoue, J.-M., Wielgosz-Collin, G., Houssay, C., Kornprobst, J.-M., Debitus, C., et al. (2004) Glycolipids from marine sponges□: monoglycosylceramides and alkyldiglycosylglycerols:

isolation, characterization and biological activity. *Boll. dei Musei e Degli Ist. Biol. dell'Università di Genova* **68**: 327–334.

Gloeckner, V., Lindquist, N., Schmitt, S., and Hentschel, U. (2013) *Ectyoplasia ferox*, an experimentally tractable model for vertical microbial transmission in marine sponges. *Microb. Ecol.* **65**: 462–474.

Gloeckner, V., Wehrl, M., Moitinho-Silva, L., Gernert, C., Hentschel, U., Schupp, P., et al. (2014) The HMA-LMA dichotomy revisited: an electron microscopical survey of 56 sponge species. *Biol. Bull.* **227**: 78–88.

Grissa, I., Vergnaud, G., and Pourcel, C. (2008) CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **36**: 52–57.

Habyarimana, F., Al-Khodor, S., Kalia, A., Graham, J.E., Price, C.T., Garcia, M.T., and Kwaik, Y.A. (2008) Role for the Ankyrin eukaryotic-like genes of *Legionella pneumophila* in parasitism of protozoan hosts and human macrophages. *Environ. Microbiol.* **10**: 1460–1474.

Haft, D.H., Selengut, J.D., and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**: 371–373.

Hansen, I. V., Weeks, J.M., and Depledge, M.H. (1995) Accumulation of copper, zinc, cadmium and chromium by the marine sponge *Halichondria panicea* Pallas and the implications for biomonitoring. *Mar. Pollut. Bull.* **31**: 133–138.

Hentschel, U., Fieseler, L., Wehrl, M., Gernert, C., Steinert, M., Hacker, J., and Horn, M. (2003) Microbial diversity of marine sponges. In, *Sponges (Porifera)*. Springer, pp. 59–88.

Hentschel, U., Hopke, J., Horn, M., Friedrich, A.B., Wagner, M., Hacker, J., and Moore, B.S. (2002) Molecular evidence for a uniform microbial community in sponges from different oceans. *Appl. Environ. Microbiol.* **68**: 4431–4440.

Hentschel, U., Piel, J., Degnan, S.M., and Taylor, M.W. (2012) Genomic insights into the marine sponge microbiome. *Nat. Rev. Microbiol.* **10**: 641–654.

Hooper, J.A. and Van Soest, R.M. (2002) Systema Porifera. A guide to the classification of sponges Hooper,J.A. and Van Soest,R.M. (eds) Springer US.

Horn, H., Slaby, B., Jahn, M., Bayer, K., Moitinho-Silva, L., Förster, F., et al. (2016) An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. *Front. Microbiol.* **7**: 1751. doi: 10.3389/fmicb.2016.01751.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., et al. (2016) A new view of the tree of life. *Nat. Microbiol.* **1**: 16048. doi: 10.1038/nmicrobiol.2016.48.

Kamke, J., Rinke, C., Schwientek, P., Mavromatis, K., Ivanova, N., Sczyrba, A., et al. (2014) The candidate phylum *Poribacteria* by single-cell genomics: new insights into phylogeny, cell-compartmentation, eukaryote-like repeat proteins, and other genomic features. *PLoS One* **9**: e87353. doi:10.1371/journal.pone.0087353.

Kamke, J., Sczyrba, A., Ivanova, N., Schwientek, P., Rinke, C., Mavromatis, K., et al. (2013) Single-cell genomics reveals complex carbohydrate degradation patterns in poribacterial symbionts of marine sponges. *ISME J.* **7**: 2287–300.

Kamke, J., Taylor, M.W., and Schmitt, S. (2010) Activity profiles for marine sponge-associated bacteria obtained by 16S rRNA vs 16S rRNA gene comparisons. *ISME J.* **4**: 498–508.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**: 277–280.

Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **428**: 726–731.

Keeling, P.J. and Palmer, J.D. (2008) Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*

9: 605–618.

Kvint, K., Nachin, L., Diez, A., and Nyström, T. (2003) The bacterial universal stress protein: function and regulation. *Curr. Opin. Microbiol.* **6**: 140–145.

Lackner, G., Peters, E.E., Helfrich, E.J.N., and Piel, J. (2017) Insights into the lifestyle of uncultured bacterial natural product factories associated with marine sponges. *Proc. Natl. Acad. Sci. USA* **114**: E347–E356.

Liu, M., Fan, L., Zhong, L., Kjelleberg, S., and Thomas, T. (2012) Metaproteogenomic analysis of a community of sponge symbionts. *ISME J.* **6**: 1515–1525.

Liu, M.Y., Kjelleberg, S., and Thomas, T. (2011) Functional genomic analysis of an uncultured δ-proteobacterium in the sponge *Cymbastela concentrica*. *ISME J.* **5**: 427–435.

López-Legentil, S., Song, B., McMurray, S.E., and Pawlik, J.R. (2008) Bleaching and stress in coral reef ecosystems: hsp70 expression by the giant barrel sponge *Xestospongia muta*. *Mol. Ecol.* **17**: 1840–9.

Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., et al. (1998) Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**: 554–568.

Luter, H.M., Whalan, S., and Webster, N.S. (2012) Thermal and sedimentation stress are unlikely causes of brown spot syndrome in the coral reef sponge, *Ianthella basta*. *PLoS One* **7**: e39779. doi:10.1371/journal.pone.0039779.

Mackie, W. and Preston, R.D. (1968) The occurrence of mannan microfibrils in the green algae *Codium fragile* and *Acetabularia crenulata*. *Planta* **253**: 249–253.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., et al. (2011) Evolution and classification of the CRISPR–Cas systems. *Nat. Rev. Microbiol.* **9**: 467–477.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., et al. (2015) An

updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**: 722–736.

McCord, J.M. and Fridovich, I. (1969) Superoxide dismutase. An enzymic function for erythrocuprein (hemocuprein). *J. Biol. Chem.* **244**: 6049–6055.

Moitinho-Silva, L., Steinert, G., Nielsen, S., Hardoim, C.C.P., Wu, Y.-C., McCormack, G.P., et al. (2017) Predicting the HMA-LMA status in marine sponges by machine learning. *Front. Microbiol.* **8**: 752. doi: 10.3389/fmicb.2017.00752.

Montalvo, N.F. and Hill, R.T. (2011) Sponge-associated bacteria are strictly maintained in two closely related but geographically distant sponge hosts. *Appl. Environ. Microbiol.* **77**: 7207–7216.

Newman, D.J. and Cragg, G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* **75**: 311–335.

Nguyen, M., Ekstrom, A., Li, X., and Yin, Y. (2015) HGT-finder: a new tool for horizontal gene transfer finding and application to *Aspergillus* genomes. *Toxins* **7**: 4035–4053.

Nyström, T. and Neidhardt, F.C. (1994) Expression and role of the universal stress protein, UspA, of *Escherichia coli* during growth arrest. *Mol Microbiol* **11**: 537–544.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., et al. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**: 206–214.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**: 1043–1055.

Parks, D.H., Tyson, G.W., Hugenholtz, P., and Beiko, R.G. (2014) STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* **30**: 3123–3124.

Pawlik, J.R. (2011) The chemical ecology of sponges on Caribbean reefs: natural products shape

natural systems. *Bioscience* **61**: 888–898.

Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Piel, J. (2009) Metabolites from symbiotic bacteria. *Nat. Prod. Rep.* **26**: 338–362.

Piel, J., Hui, D., Wen, G., Butzke, D., Platzer, M., Fusetani, N., and Matsunaga, S. (2004) Antitumor polyketide biosynthesis by an uncultivated bacterial symbiont of the marine sponge *Theonella swinhoei*. *Proc. Natl. Acad. Sci. USA* **101**: 16222–16227.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glöckner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**: 7188–96.

Reynolds, D. and Thomas, T. (2016) Evolution and function of eukaryotic-like proteins from sponge symbionts. *Mol. Ecol.* **25**: 5242–5253.

Ribes, M., Calvo, E., Movilla, J., Logares, R., Coma, R., and Pelejero, C. (2016) Restructuring of the sponge microbiome favors tolerance to ocean acidification. *Environ. Microbiol. Rep.* **8**: 536–544.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Sala, G. Della, Hochmuth, T., Teta, R., Costantino, V., and Mangoni, A. (2014) Polyketide synthases in the microbiome of the marine sponge *Plakortis halichondrioides*: a metagenomic update. *Mar. Drugs* **12**: 5425–5440.

Schmitt, S., Tsai, P., Bell, J., Fromont, J., Ilan, M., Lindquist, N., et al. (2012) Assessing the complex sponge microbiota: core, variable and species-specific bacterial communities in marine

sponges. *ISME J.* **6**: 564–576.

Siegl, A., Kamke, J., Hochmuth, T., Piel, J., Richter, M., Liang, C., et al. (2011) Single-cell genomics reveals the lifestyle of *Poribacteria*, a candidate phylum symbiotically associated with marine sponges. *ISME J.* **5**: 61–70.

Simister, R., Deines, P., Botté, E.S., Webster, N.S., and Taylor, M.W. (2012a) Sponge-specific clusters revisited: a comprehensive phylogeny of sponge-associated microorganisms. *Environ. Microbiol.* **14**: 517–524.

Simister, R., Taylor, M.W., Rogers, K.M., Schupp, P.J., and Deines, P. (2013) Temporal molecular and isotopic analysis of active bacterial communities in two New Zealand sponges. *FEMS Microbiol. Ecol.* **85**: 195–205.

Simister, R., Taylor, M.W., Tsai, P., and Webster, N. (2012b) Sponge-microbe associations survive high nutrients and temperatures. *PLoS One* **7**: e52220. doi:10.1371/journal.pone.0052220.

Slaby, B.M., Hackl, T., Horn, H., Bayer, K., and Hentschel, U. (2017) Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *ISME J.* 1–15. doi:10.1038/ismej.2017.101.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E. V, et al. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: 41.

Taylor, M.W., Radax, R., Steger, D., and Wagner, M. (2007a) Sponge-associated microorganisms: evolution, ecology, and biotechnological potential. *Microbiol. Mol. Biol. Rev.* **71**: 295–347.

Taylor, M.W., Thacker, R.W., and Hentschel, U. (2007b) Evolutionary insights from sponges. *Science* **316**: 1854–1855.

Taylor, M.W., Tsai, P., Simister, R.L., Deines, P., Botte, E., Ericson, G., et al. (2013) "Sponge-specific" bacteria are widespread (but rare) in diverse marine environments. *ISME J.* **7**: 438–443.

Thomas, T., Moitinho-Silva, L., Lurgi, M., Björk, J.R., Easson, C., Astudillo-García, C., et al. (2016) Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun.* **7**: 11870. doi:10.1038/ncomms11870.

Thomas, T., Rusch, D., DeMaere, M.Z., Yung, P.Y., Lewis, M., Halpern, A., et al. (2010) Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J.* **4**: 1557–1567.

Tian, R.-M., Zhang, W., Cai, L., Wong, Y.-H., Ding, W., and Qian, P.-Y. (2017) Genome reduction and microbe-host interactions drive adaptation of a sulfur-oxidizing bacterium associated with a cold seep sponge. *mSystems* **2**: e00184-16.

Tian, R.M., Wang, Y., Bougouffa, S., Gao, Z.M., Cai, L., Bajic, V., and Qian, P.Y. (2014) Genomic analysis reveals versatile heterotrophic capacity of a potentially symbiotic sulfur-oxidizing bacterium in sponge. *Environ. Microbiol.* **16**: 3548–3561.

Vasu, K. and Nagaraja, V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.* **77**: 53–72.

Wagner, M. and Horn, M. (2006) The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae* and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17**: 241–249.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H.U., Bruccoleri, R., et al. (2015) AntiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **43**: 237–243.

Webster, N.S., Botté, E.S., Soo, R.M., and Whalan, S. (2011) The larval sponge holobiont exhibits high thermal tolerance. *Environ. Microbiol. Rep.* **3**: 756–762.

Webster, N.S. and Thomas, T. (2016) The sponge hologenome. *MBio* **7**: e00135–16. doi:10.1128/mBio.00135–16.

Webster, N.S., Webb, R.I., Ridd, M.J., Hill, R.T., and Negri, A.P. (2001) The effects of copper on the microbial community of a coral reef sponge. *Environ. Microbiol.* **3**: 19–31.

Wheeler, T.J. and Eddy, S.R. (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**: 2487–2489.

Wiedenbeck, J. and Cohan, F.M. (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol. Rev.* **35**: 957–976.

Wilson, M.C., Mori, T., Rückert, C., Uria, A.R., Helf, M.J., Takada, K., et al. (2014) An environmental bacterial taxon with a large and distinct metabolic repertoire. *Nature* **506**: 58–62.

Wortham, B.W., Patel, C.N., and Oliveira, M.A. (2007) Polyamines in bacteria: pleiotropic effects yet specific mechanisms. In, Skurnik,M., Bengoechea,J.A., and Granfors,K. (eds), *The genus Yersinia*. Springer US, pp. 106–115.

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011) WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12**: 444.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., et al. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* **12**: 635–645.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012) DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**: 445–451.

Youssef, N.H., Farag, I.F., Rinke, C., Hallam, S.J., Woyke, T., and Elshahed, M.S. (2015) *In silico* analysis of the metabolic potential and niche specialization of candidate phylum "*Latescibacteria*" (WS3). *PLoS One* **10**: e0127499.

Zhang, F., Blasiak, L.C., Karolin, J.O., Powell, R.J., Geddes, C.D., and Hill, R.T. (2015) Phosphorus sequestration in the form of polyphosphate by microbial symbionts in marine sponges. *Proc.*

Natl. Acad. Sci. USA **112**: 4381–4386.

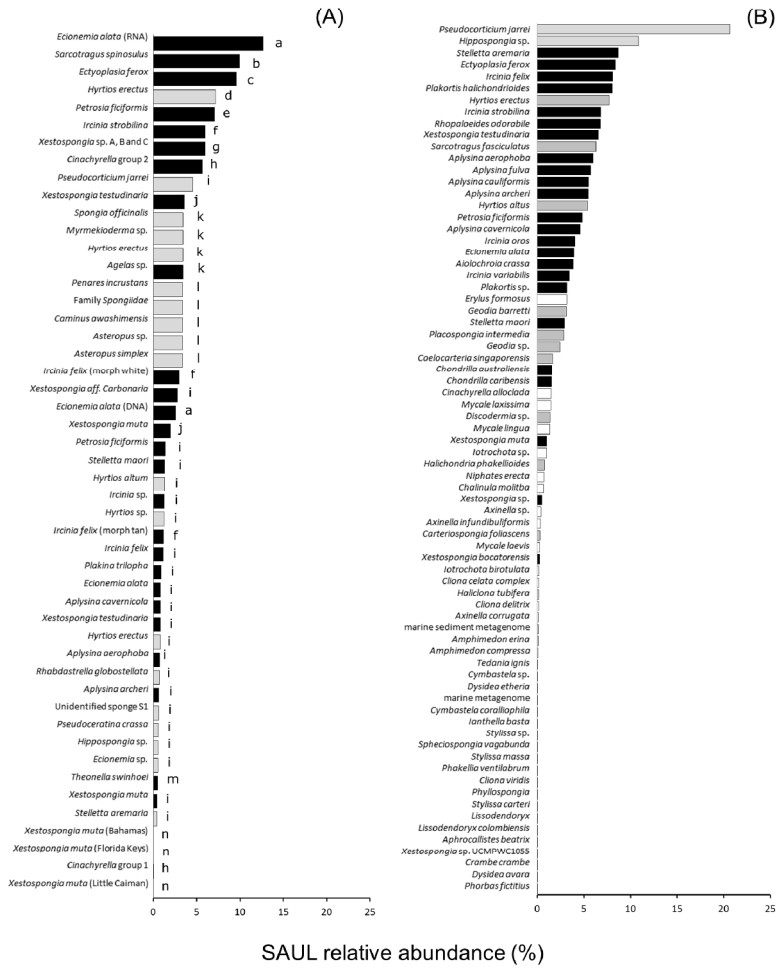

Figure 1: Relative abundance (percentage) of SAUL- or PAUC34f-affiliated sequences in 16S rRNA gene-based studies of marine sponge bacterial communities. (A) Prevalence of SAUL in different sponge species across 14 sponge microbiota studies. Letters beside bars represent sponge species belonging to the same study defined in Table S1. (B) Prevalence of SAUL in 72 of the 81 sponge species reported in the recent Sponge Microbiome Project study (Thomas et al., 2016). Where known, sponge species are classified as high (black bars) or low microbial abundance (white bars), according mainly to Gloeckner et al. (2014) and Moitinho-Silva et al. (2017)

190x275mm (300 x 300 DPI)

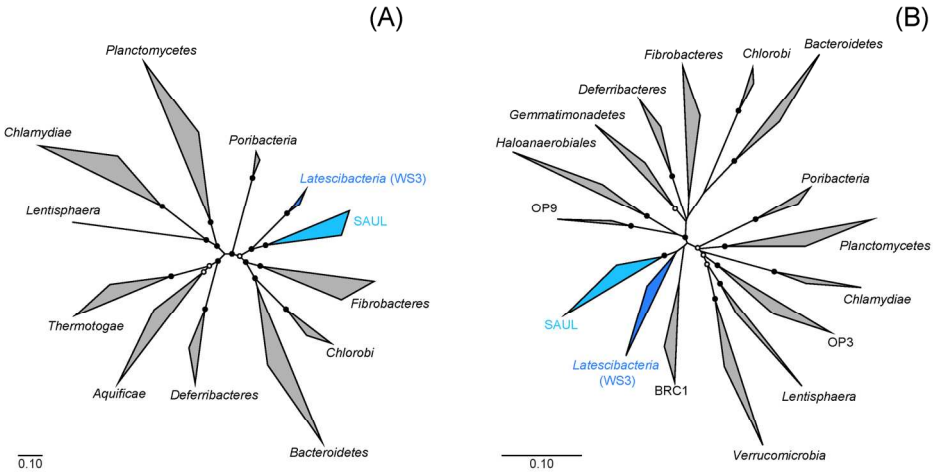

Figure 2: Phylogeny showing position of the SAUL clade relative to other bacterial phyla. (A) Phylogenomic analysis based on a data set of up to 37 concatenated marker proteins. (B) 16S rRNA gene sequence-based maximum-likelihood analysis of SAUL and its closest relatives. Bootstrap support (100 resamplings for (A), 500 for (B)) is shown on tree nodes where support is either ≥90% (filled circles) or ≥75% (open circles). Scale bars represent 10% sequence divergence.

254x190mm (300 x 300 DPI)

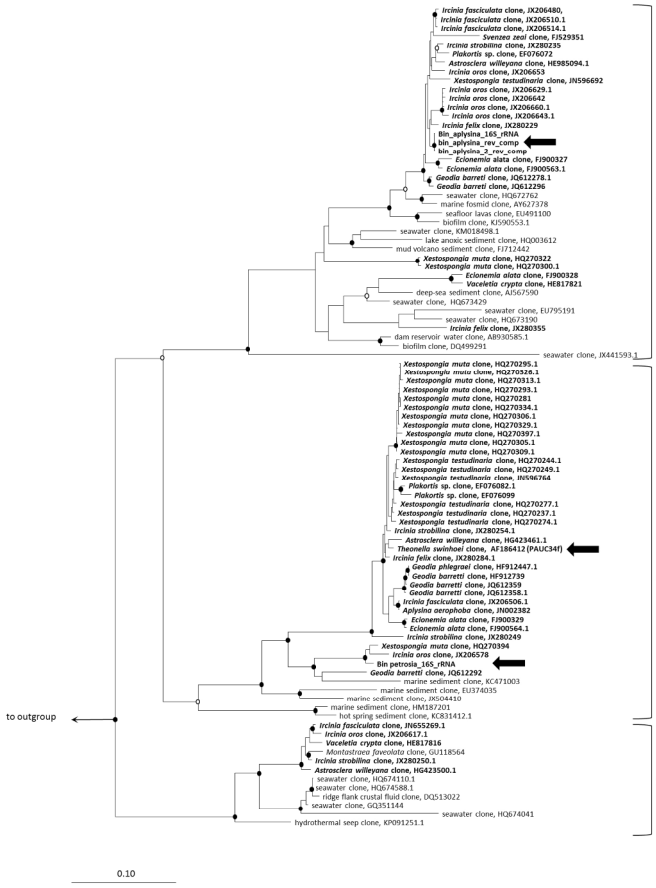

Figure 3. 16S rRNA gene-based maximum likelihood phylogenetic analysis showing the internal architecture of the SAUL clade. Bold sequences denote sequences derived from marine sponges; black arrows indicate sequences obtained from SAUL bins (bin_aplysina, bin_aplysina_2 and bin_petrosia) as well as the first SAUL/PAUC34f sequence identified (AF186412). Details are the same as those provided for Figure 2A.

254x338mm (300 x 300 DPI)

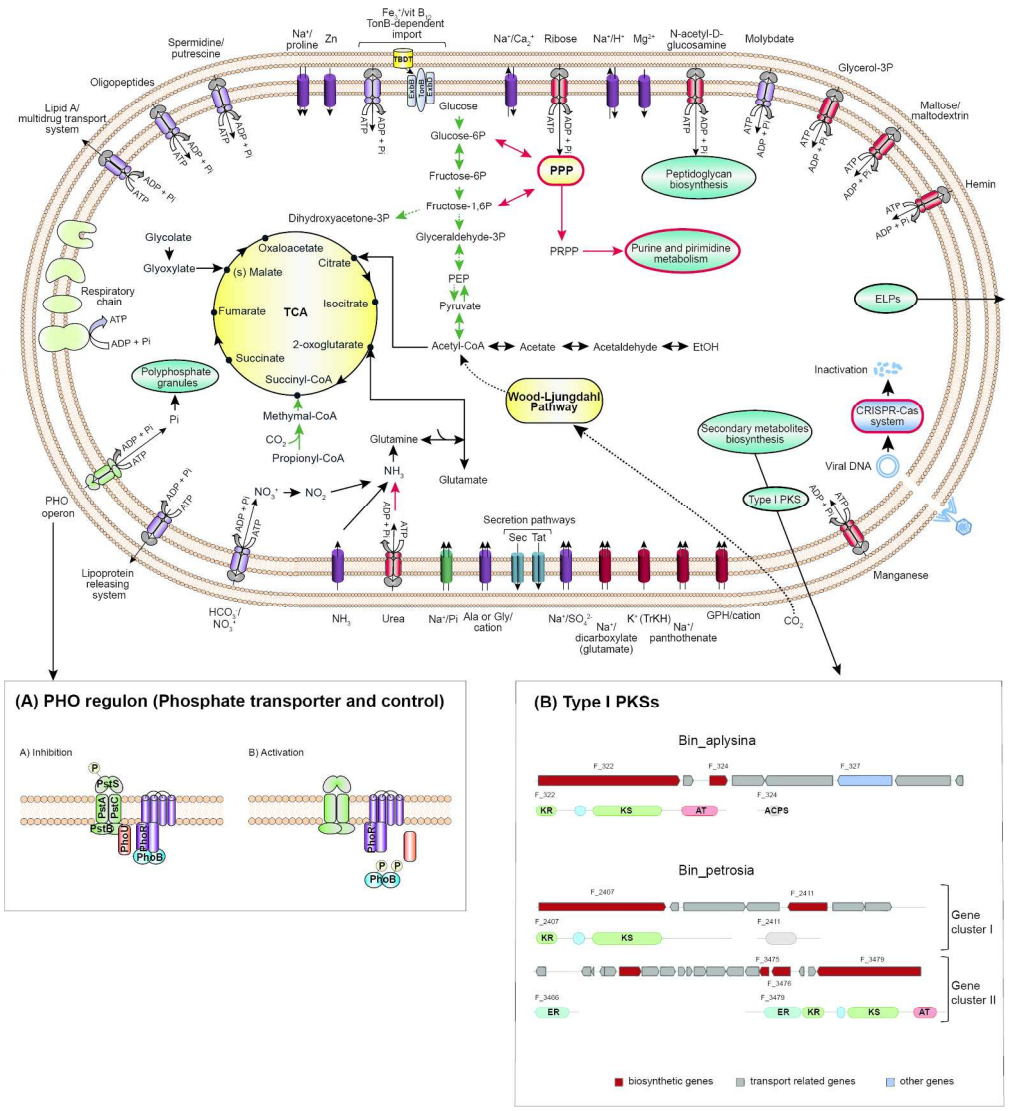

Figure 4. Schematic overview showing genomic potential of a SAUL cell. Black lines indicate functions identified in both metagenome bins. Red and green lines/transporters indicate functions only present in either bin_aplysina or bin_petrosia, respectively. Dashed lines indicate that some enzymes of the pathway were not identified in the bins (see Supporting Text for details). (A) Model of phosphate regulon signal transduction; (B) Features of the Type I polyketide synthase identified in bin_aplysina and bin_petrosia. Abbreviations of PKS domains as follows: KR, ketoreductase; KS, ketosynthase; AT, acyltransferase; ACPS, acyl carrier protein; ER, Enoyl reductase.

207x228mm (300 x 300 DPI)

Table 1. General information about SAUL metagenome bins.

	bin_aplysina	bin_petrosia
Assembly size (bp)	5,661,056	4,100,466
Estimated genome completeness (%)	90.38	86.54
Estimated total genome size (bp)	6,263,616	4,738,232
Estimated genome contamination (%)	3.85	8.65
Number of contigs	632	349
GC content (%)	58.5	59.5
n50	11744	13499
SEED subsystems	285	265
Protein CDs		
Number	4887	3675
%	99.09	99.03
Protein in subsystem (total)/SEED functions	1342	962
Non-hypothetical	1294	933
Hypothetical	48	29
Protein not in subsystem (total)	3545	2713
Non-hypothetical	1276	1034
Hypothetical	2269	1679
Protein coding genes with function prediction		
Number	2570	1967
% of total number of genes	52.11	53.00
Protein coding genes without function prediction		
Number	2317	1708
% of total number of genes	46.98	46.03
rRNAs (5S rRNA, 16S rRNA, 23S rRNA)		
Number	3 (1, 1, 1)	3 (1, 1, 1)
% of total number of genes	0.06	0.08
tRNAs		
Number	42	33
% of total number of genes	0.85	0.89
COGs (unique)	1349	1231
COGs (total)	3174	2427